

Understanding Crime Trends in Los Angeles through 911 Call Data Analysis

Md Tahidul Islam, Abhinav Adhikari, Sathish Reddy Kallu, Vikas Reddy

December 13, 2023

Abstract

This dataset offers valuable insights into criminal activity patterns, providing crucial information for law enforcement and policymakers. While the dataset represents reported incidents, incorporating broader population data would enhance our understanding. The analysis aims to uncover key trends, emphasizing the need for future research to integrate census data for comprehensive insights. The dataset covers reported crime incidents in Los Angeles from 2020 onwards. Notably, the data originates from transcriptions of original reports, potentially introducing inaccuracies. Instances with missing location information are marked as (0°, 0°), and addresses are truncated for privacy. It is essential to consider the source data's integrity. This analysis strives to extract meaningful insights, but it solely focuses on reported incidents. Future studies could benefit from integrating additional data sources, such as census and demographic information, for a broader perspective.

1 Introduction

Los Angeles, a city renowned for its diversity and urban dynamics, faces a complex array of reported crimes. Understanding criminal activity patterns and dynamics is essential for law enforcement agencies and policymakers. This analysis explores a meticulous dataset of crime incidents to uncover crucial insights that can inform decision-making and public safety measures.

1.1 Motivation

This dataset was selected to shed light on prevailing trends and patterns of criminal incidents in Los Angeles. By analyzing reported incidents, we aim to contribute to a broader understanding of the factors influencing crime in the city. However, it is important to note that this dataset only captures reported incidents. Integrating demographic and census data could provide a more nuanced understanding of the underlying socioeconomic factors that may contribute to these incidents.

1.2 Data Source

The dataset originates from official records of reported incidents in Los Angeles, encompassing incidents documented from 2020 onwards. It is important to note that the data is derived from transcriptions of original reports, which introduces the possibility of minor discrepancies. Instances with missing location information are represented by coordinates (0°, 0°), and addresses are truncated to safeguard privacy. The reliability of the source data must be considered in the interpretation of results.

1.3 Significance of the Problem

Understanding the dynamics of reported incidents in Los Angeles is paramount for various stakeholders. Law enforcement agencies can utilize this information to allocate resources efficiently and deploy preventive measures in areas with higher incident rates. Policymakers can glean insights to formulate targeted policies that address the underlying causes of criminal activity.

1.4 Structure of the Analysis

This analysis is structured into five key sections:

1. Introduction: Provides an overview of the motivation, data source, and significance of the analysis.

2. Data Description: Offers details about the dataset, including its source, quality, structure, variables, and utility. Additionally, this section outlines how the data was cleaned and prepared.
3. Visualization and Analysis: Presents various data visualizations and analyses, including: Crime count by area Most frequent times of crime occurrence Top crime descriptions Crime by age range Comparison of male and female victims over the years Crime incidents by the day of the week and month Breakdown of crime by victim race Exploring LA Crime through Leaflet
4. Conclusion and Future Directions: Summarizes the key findings of the analysis and outlines potential future directions for more in-depth research, including the need to integrate additional data sources for a more comprehensive perspective.

In conclusion, this project draft serves as a preliminary exploration of crime incidents in Los Angeles, shedding light on important trends and patterns. The findings presented here can serve as a foundation for future research and policy development aimed at enhancing the safety and security of Los Angeles residents.

2 Understanding and Cleaning The Data

2.1 What is the Dataset primarily about?

The dataset is related to crimes, specifically those committed in city of Los Angeles. It includes various attributes such as the year of the crime, time of occurrence, location (area), a description of the crime, latitude and longitude coordinates, the month of the crime, the age of the victim, and the reporting district number.

2.2 How Was the Data Processed?

Column Name Standardization and Formatting: Dots in column names were replaced with underscores for consistency. All column names were converted to title case (i.e., the first letter of each word capitalized).

Date and Time Formatting: The "12:00:00 AM" portion was removed from the date column, leaving only the date. Military time format in the 'Time_occ' column was standardized to HH:MM format, with leading zeros added as needed.

Weekday, Month, and Year Extraction: New columns were created to extract the weekday, month, and year of occurrence from the 'Date_occ' column. The lubridate package was used for this operation.

Victim Descent Standardization: Abbreviated codes in the 'Vict_descent' column were replaced with their corresponding full names using a predefined mapping.

2.3 What Was Done to Clean the Data? Detailed Explanation

Step 1: Making Column Names Clearer

At first, we work on the names of the columns in our dataset. We do two things:

- a. Changing Dots to Underscores: If there are any dots in the column names, we swap them with underscores. This helps make the names more understandable.
- b. Capitalizing Words: We also make sure the column names start with capital letters. This makes them look neater and easier to read.

Step 2: Fixing Time Information

Next, we deal with times in our data. We want to get rid of the "12:00:00 AM" part. We go through all the columns and replace this time with nothing. This way, we're left with just the date.

Step 3: Using Standard Time Format

We want all our times to look the same way - with four numbers. We make sure they're written like this: "HH:MM". This helps keep things consistent and clear.

Step 4: Adding More Date Details

We want to know more about the dates when things happened. So, we create three new pieces of information:

- Day of the Week: We find out which day of the week each event occurred on and write it down.
- Month: We note down which month each event happened in.
- Year: We also record the year when each event took place.

Step 5: Understanding Victim Details

We want to better understand who the victims are. We make a list that connects short codes to full names, based on the data dictionary from <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8> (<https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>). Then, we use the `dplyr::mutate()` function to replace the short codes with the full names in the `Vict_descent` column.

Step 6: Saving Our Cleaned Data

Lastly, we save our cleaned data into a file. This way, we can use it for later without having to do all these steps again. We name the file “cleaned_test1.csv”. We also make sure not to include extra row numbers in the file to keep it clean.

Now, the cleaned data can be used for further analysis or reporting purposes.

3 What does the Data Say?

The analysis of the dataset provides valuable insights into various aspects of crime in the region. Through a combination of visualizations and data exploration, we gain a deeper understanding of patterns, trends, and demographics related to criminal activities.

3.1 Crime Count by Area

Description: This bar chart displays the count of crimes in different areas. Each bar represents an area, and the height of the bar indicates the number of reported crimes in that area.

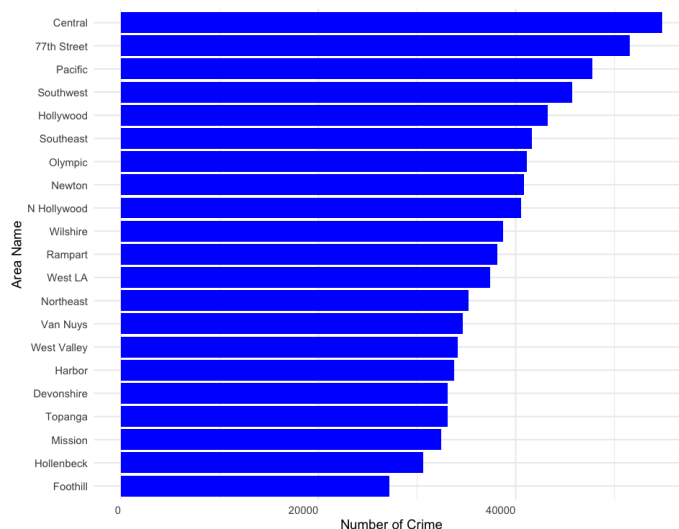


Figure 1: Bar chart showing the number of incidents for each crime type.

How the code works:

The code first loads the necessary libraries, including `dplyr`, `ggplot2`, and `knitr`. Then, it groups the crime data by area and calculates the number of crimes in each area. Next, it reorders the areas by the number of crimes in descending order. Finally, it creates a bar plot showing the count of crimes by area, with the areas with the highest crime counts listed first. From the code, we get to know that the area with the highest crime count is central with 54841 crime reported and foothill with the least crime reported of 27225.

3.2. Crime Count by Time Range

Description: This line graph illustrates the count of crimes that occurred within specific time ranges. The x-axis represents the time ranges, while the y-axis shows the corresponding number of reported crimes.

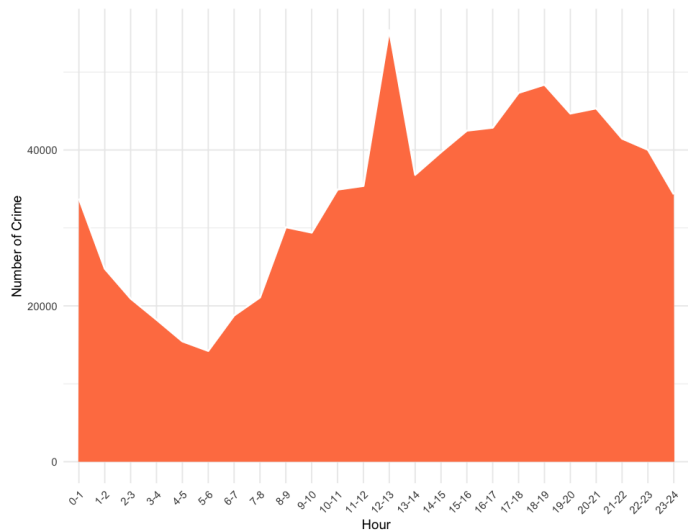


Figure 2: Line graph showing the number of crimes by time range.

How the code works:

The code first extracts the hour component from the Time_occ column. Then, it groups the data by the hour and calculates the number of crimes that occurred within each hour. Next, it creates a new column called Time_Range, which contains the time ranges in the format HH-HH+1. For example, the time range for 1 AM to 2 AM is 1-2. Finally, the code arranges the data by the hour in ascending order and reorders the Time_Range column based on the hour.

The last part of the code creates a line graph showing the count of crimes by time range. The geom_area() function is used to create a shaded area graph, with the color and fill arguments set to white and coral, respectively. The labs() function is used to add labels to the plot, and the theme() function is used to customize the appearance of the plot.

Interpretation:

From the code block we were able to isolate that the crimes most frequently reported was from 12-1 pm and the least reported was 5-6 am. And by the time passing the crime reported also grows.

3.3. Seven Most Frequent Crime Descriptions

Description: This vertical bar chart highlights the seven most common crime descriptions. Each bar represents a specific crime description, and its height indicates the frequency of occurrences.

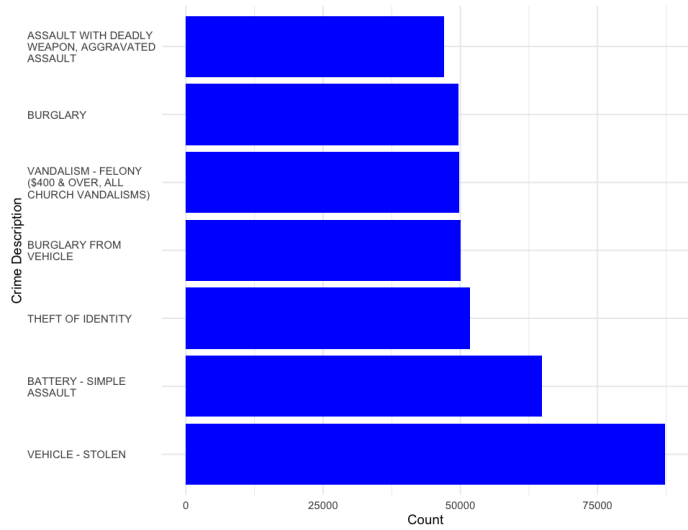


Figure3 : Vertical Barchart to show most frequent crimes

How the code works: It groups the crime data by the crime description and calculates the number of occurrences for each crime description. Next, it arranges the data in descending order by the number of occurrences and selects the top seven crime descriptions.

To wrap the text labels on the y-axis, the code uses the `str_wrap()` function from the `stringr` library. This function splits the text labels into multiple lines, with each line having a maximum width of 20 characters. The `width` argument can be adjusted as needed.

Finally, the code creates a vertical bar chart showing the frequency of occurrences for the top seven crime descriptions. The `geom_bar()` function is used to create the bar chart, with the `fill` argument set to blue. The `labs()` function is used to add labels to the plot, and the `theme()` function is used to customize the appearance of the plot.

3.4. Crime Curve by Age Range

Description: This line plot depicts the variation in crime counts across different age ranges. The x-axis represents the age ranges, and the y-axis shows the corresponding count of reported crimes.

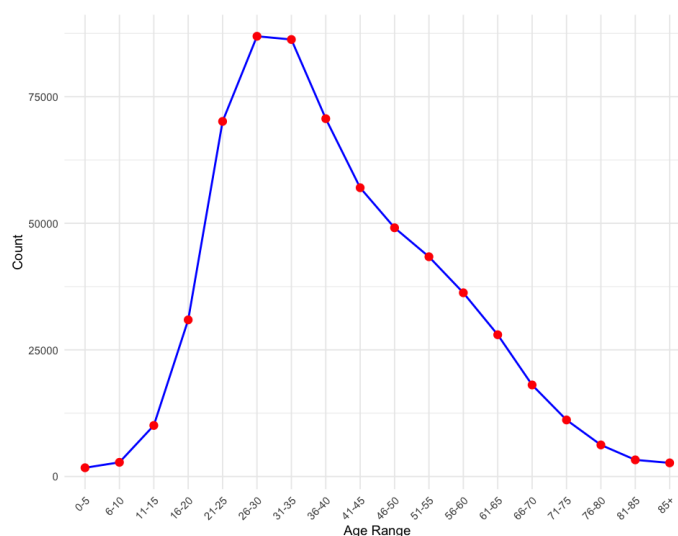


Figure 4 : Line Plot to depict various crime counts across different age range.

How the code works:

The code first loads the necessary libraries, including `dplyr`, `ggplot2`, and `knitr`. Then, it converts any non-numeric values in the `Vict_age` column to `NA`. Next, it creates a new column for age ranges, excluding `NA` values. The code uses the `cut()` function to create the age ranges, with the following breakpoints:

`c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, Inf)` The corresponding age range labels are:

`c("0-5", "6-10", "11-15", "16-20", "21-25", "26-30", "31-35", "36-40", "41-45", "46-50", "51-55", "56-60", "61-65", "66-70", "71-75", "76-80", "81-85", "85+")`

Next, the code groups and summarizes the data by age range, excluding `NA` values. It then creates a line plot for the crime curve, using the `geom_line()` and `geom_point()` functions. The `labs()` function is used to add labels to the plot, and the `theme()` function is used to customize the appearance of the plot.

Interpretation:

Crime rates tend to go up as people get older, peaking in the late teens and early twenties. After that, crime rates slowly go down as people get older. This is consistent with the age-crime curve, which is a well-known finding in criminology.

There are a few possible explanations for the age-crime curve. One possibility is that young people are more likely to take risks, which can lead to crime. Another possibility is that young people are more likely to be victims of crime, which can make them more likely to commit crimes themselves.

3.5. Top Three Crimes by Age Range (20-60)

Description: This line graph focuses on the top three crimes within the age range of 20-60. Each line represents a specific crime, and the x-axis denotes the age range, while the y-axis displays the corresponding count.

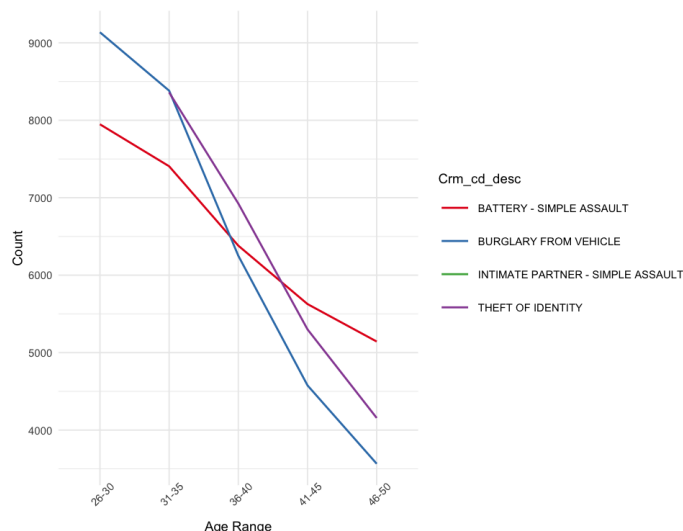


Figure 5: Line Plot displaying the highest committed crime by age group(20-60).

How the code works:

The first code chunk filters the crime data to only include people in the age range of 20-60. Then, it groups the data by age range and crime description and counts the number of crimes in each group. Finally, it selects the top three most common crimes for each age range.

The second code chunk creates a line graph showing the top three crimes by age range. The `geom_line()` function is used to create the line graph, with the color argument set to the crime description. The `labs()` function is used to add labels to the plot, and the `theme()` function is used to customize the appearance of the plot.

3.6. Comparison of Male and Female Victims

Description: This side-by-side histogram offers a visual comparison between male and female victims of crimes. The x-axis represents the years, and the y-axis shows the count of reported crimes.

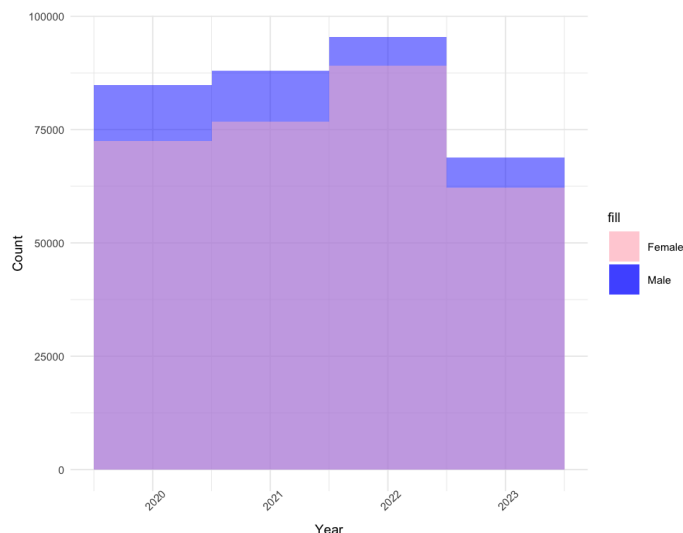


Figure 6 : Comparison of Male and Female Victims from Jan 2020 to Sep 2023

How the code works: This code generates a side-by-side histogram comparing male and female victims of crimes over the specified years. It utilizes `dplyr` for data manipulation, `ggplot2` for visualization, and `knitr` for dynamic report generation. The data is grouped by gender and year, and counts are calculated for male and female victims. The resulting counts are used to create the histogram, where blue bars represent male victims and pink bars represent female victims. The plot offers a visual comparison of reported crimes between the two genders over the specified years.

Interpretation: It can be observed that there were 337,050 reported crimes involving male victims and 300,602 reported crimes involving female victims. This indicates a higher count of reported crimes involving male victims compared to female victims. The higher number of reported crimes involving male victims may be influenced by factors like occupational risks, risk-taking behavior or demographic distribution.

3.7. Average Crimes by Day of the Week

Description: This bar chart depicts the average number of crimes that occur in different months. The x-axis represents the months, and the y-axis shows the corresponding average count.

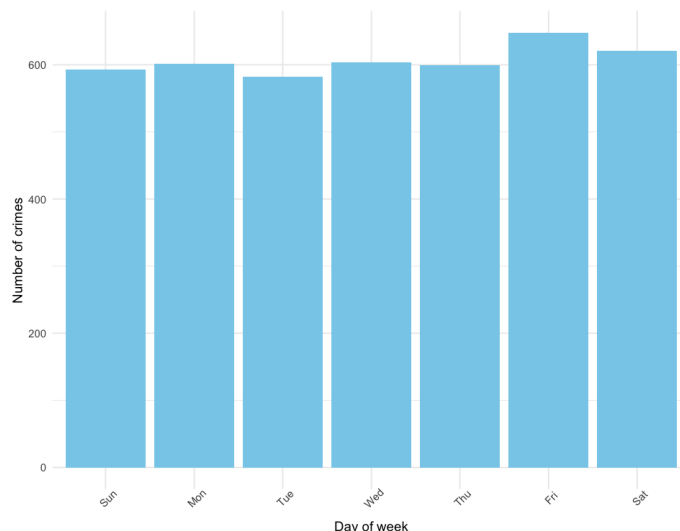


Figure 7: Bar Chart to show crimes reported to 911 on average per day of a week.

How the code works: The code first groups the data frame by the day of the week on which each crime occurred. This creates a new data frame with one row for each day of the week, and one column containing the number of crimes that occurred on that day of the week.

The code then uses the ggplot2 package to create a bar chart. The x-axis of the bar chart represents the day of the week, and the y-axis represents the number of crimes. The bars of the chart are colored skyblue, and the chart has a caption of "Average Crimes by day of the week".

Interpretation:

The bar chart shows that the average number of crimes is highest on Saturdays and Sundays, and lowest on Tuesdays. This suggests that there are more crimes reported on weekends compared to weekdays.

This finding is likely due to the fact that people are more active and social on weekends, which makes them more vulnerable to crimes. For example, people are more likely to be out and about in public places on weekends, and they may be more likely to consume alcohol, which can impair their judgment and make them more likely to become victims of crime.

3.8. Average Crimes by Month

Description: This bar chart depicts the average number of crimes that occur in different months. The x-axis represents the months, and the y-axis shows the corresponding average count.

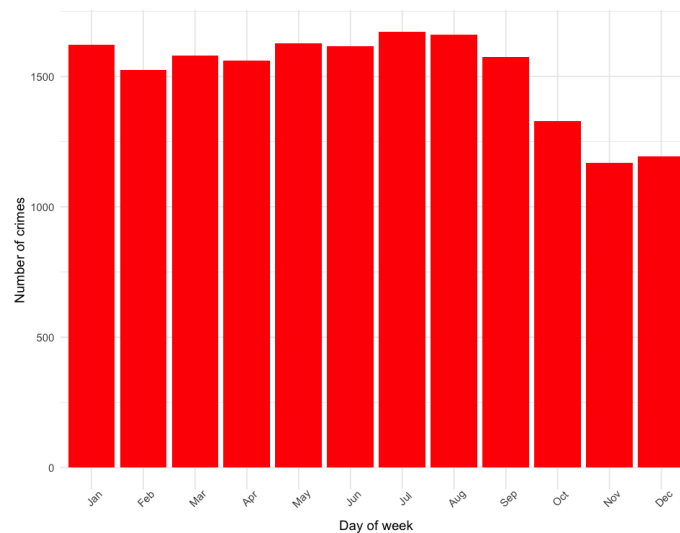


Figure 8 : Bar chart displaying number of crimes reported per month.

How the code works: The code to create a bar chart depicting the average number of crimes that occur in different months works by first grouping the crime data by month and counting the number of crimes for each month. Then, it creates a new plot object using the `ggplot()` function and specifies the aesthetic mappings for the plot, which in this case are the x-axis (month) and the y-axis (number of crimes). Next, it adds a bar chart to the plot using the `geom_bar()` function and specifies the fill color of the bars and the statistical function to be used to calculate the values of the bars. Finally, it adds labels and a caption to the plot using the `labs()` function and prints the plot.

Interpretation: The bar chart shows that crimes are more common in the summer months (July and August) than in the winter months (November and December). This is likely due to factors such as increased outdoor activity, vacations, and school breaks.

3.9. Number of Crimes by Victim Race

Description: This bar chart provides an overview of the number of crimes reported based on the victim's race. Each bar represents a specific race, and its length indicates the count of reported crimes.

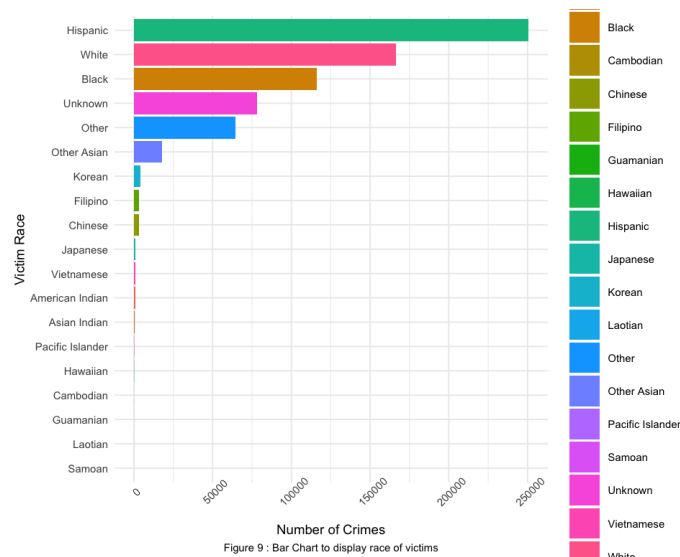


Figure 9 : Bar Chart to display race of victims

How the code Works:

It first groups the data by victim race and counts how many crimes there are for each race. Then, it removes any rows that don't have information about the victim's race. After that, it arranges the data so that the race with the most reported crimes comes first, and so on. Next, it creates a new chart using a tool called `ggplot`. It tells `ggplot` what to put on the x-axis (which shows the number of crimes), the y-axis (which shows the victim's race), and the colors of the bars for each race. It adds the bars to the chart and tells it to use the actual counts of crimes for each race. The code then puts labels on the chart to show what the axes represent and gives the chart a caption.

Interpretation: The bar chart shows that different racial groups are affected by crime in different ways. White, Black, and Hispanic people are most likely to be victims of crime.

4 Exploring Leaflet

This chapter invites you on an exciting adventure into the world of interactive data visualization using Leaflet. Together, we'll utilize code to convert crime data into a dynamic map, enabling you to delve into and unveil concealed patterns within Los Angeles' crime scene. Prepare yourself to click, explore, and unravel the secrets hidden within!

4.1 Description : Packages used and Pre-processing

Our interactive map relies on a solid groundwork of data and crucial libraries:

- a. Leaflet: This robust library constructs the map canvas, animating data points as interactive markers.
- b. dplyr: A proficient data manipulation tool, aiding in filtering and reshaping crime data according to user preferences.
- c. readr: A skillful data-reading tool seamlessly incorporating crime data, preparing it for visualization.
- d. Shiny: The architect of interactive experiences, enabling users to engage with the map and dynamically modify its content.

The preprocessing phase includes:

- i. Data refinement: Ensuring data precision and consistency by converting latitude and longitude values into a numerical format.
- ii. Area organization: Formulating a separate data frame listing available areas and their respective crime counts.
- iii. Color assignment: Utilizing the rainbow function to allocate unique, vibrant colors to each area, enhancing the visual appeal of exploration.

4.2 How is the code structured?

The code behind this interactive map is a combination of functions working together:

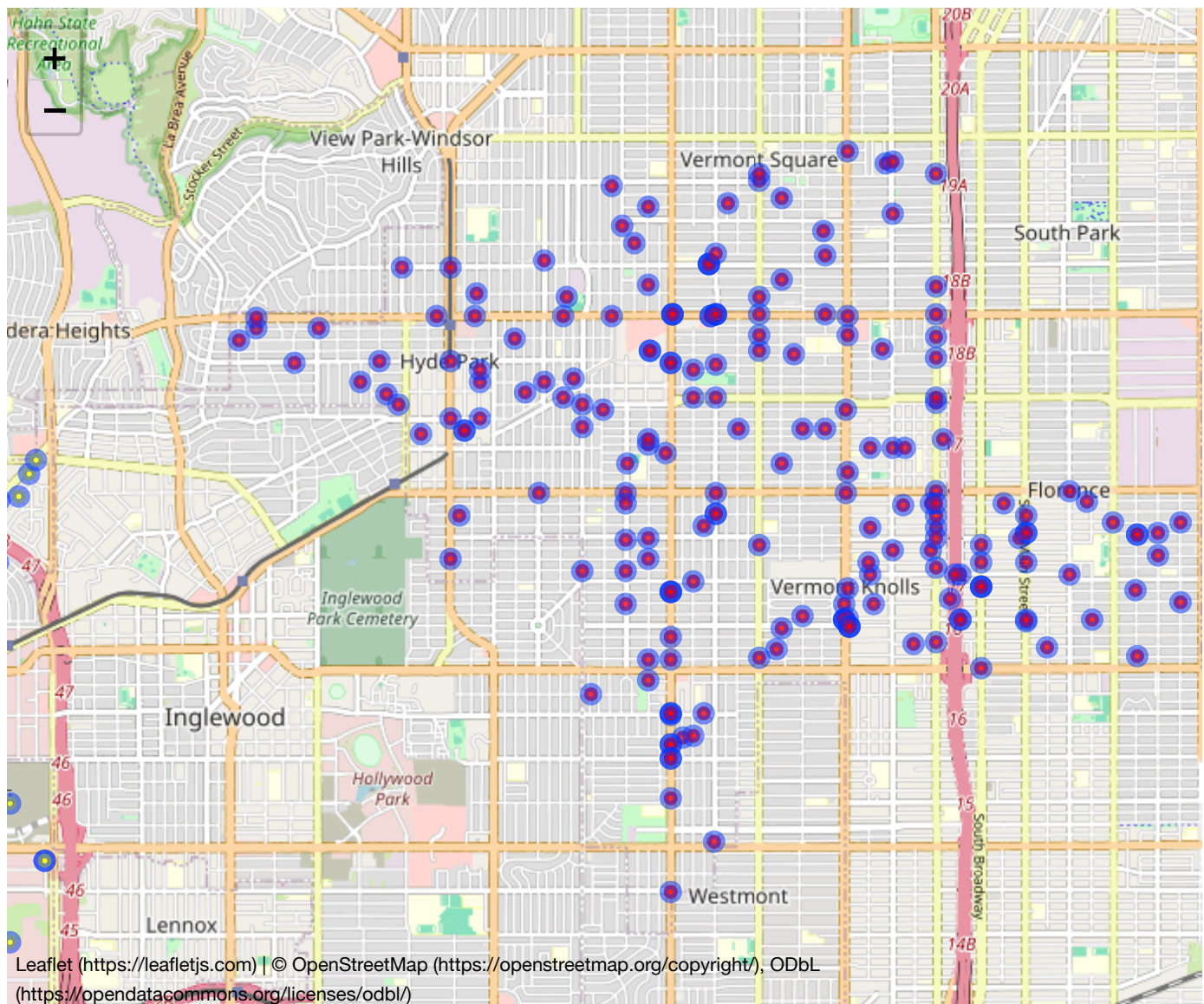
- a. selectInput: This maestro conducts the user experience, allowing you to choose areas for exploration.
- b. addTiles: It lays the foundation, adding the base map tiles for spatial context.
- c. for loop: This tireless conductor iterates through chosen areas, highlighting their unique crime patterns.
- d. addCircleMarkers: This artist paints the map with colored dots, each representing a crime within the selected area.
- e. popup: This storyteller unfolds a window into each crime, displaying its description and location.
- f. fillColor: This color coordinator assigns the appropriate color to each marker based on its corresponding area.

4.3 Your Turn to be Sherlock:

This interactive map transforms you into a data detective, empowering you to:

- a. Identify hotspots: Clusters of crime in specific areas might reveal underlying factors or potential intervention points.
- b. Track trends: Explore crime patterns over time, identifying seasonal fluctuations or area-specific trends.
- c. Compare areas: Visually compare crime rates and types across different neighborhoods, informing resource allocation strategies.

- d. Uncover hidden connections: Discover correlations between crime types and specific areas, prompting deeper investigation.



5 Conclusion

The exploration into Los Angeles' crime data has been an enriching and enlightening journey. Through understanding, cleaning, analyzing, and visualizing the data, we've unlocked valuable insights. The interactive map built with Leaflet stands as a testament to the power of data visualization in unraveling complex information. This map helps anyone become a detective, spotting problem areas, seeing patterns over time, and comparing different parts of LA.

This project serves as a testament to the potential of data-driven exploration and visualization in understanding our community and society. While our journey concludes here, the possibilities for deeper analysis, correlation discovery, and application of insights remain endless.

But this exploration has been more than just an academic exercise; it's been a humbling and eye-opening experience. Witnessing the raw data transformed into a vivid map, revealing the human stories behind each statistic, has instilled in me a profound respect for the power of data to illuminate societal challenges.

But this isn't just about playing with numbers. It's been eye-opening to see how these numbers tell real stories about people. With this map, we want to give you the power to explore your own neighborhood, find out things, and maybe even help make it safer.

Crime isn't just about numbers; it's connected to lots of things like how we live, our jobs, and the places we're in. By understanding these connections, we can make better choices about how to help our neighborhoods grow and be fairer for everyone. It's a small step, but together, we can make a big difference.

References

1. <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>
(<https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>)
2. <https://bookdown.org/yihui/rmarkdown-cookbook/hide-one.html>
(<https://bookdown.org/yihui/rmarkdown-cookbook/hide-one.html>)
3. <https://www.r-bloggers.com/2018/08/analysis-of-los-angeles-crime-with-r/> (<https://www.r-bloggers.com/2018/08/analysis-of-los-angeles-crime-with-r/>)
4. <https://www.latimes.com/california/story/2023-10-12/violent-crime-is-down-fear-is-up-why-is-la-perceived-as-dangerous> (<https://www.latimes.com/california/story/2023-10-12/violent-crime-is-down-fear-is-up-why-is-la-perceived-as-dangerous>)
5. <https://www.spolinlaw.com/criminal-defense/most-dangerous-cities-in-california/>
(<https://www.spolinlaw.com/criminal-defense/most-dangerous-cities-in-california/>)
6. <https://rstudio.github.io/leaflet/> (<https://rstudio.github.io/leaflet/>)
7. <https://dplyr.tidyverse.org/reference/> (<https://dplyr.tidyverse.org/reference/>)
8. <https://readr.tidyverse.org/> (<https://readr.tidyverse.org/>)
9. <https://shiny.posit.co/> (<https://shiny.posit.co/>)